A Bayesian approach to convolutive nonnegative matrix factorization for blind speech dereverberation

Francisco J. Ibarrola, Leandro E. Di Persia, Ruben D. Spies

PII: S0165-1684(18)30151-8
DOI: 10.1016/j.sigpro.2018.04.024
Reference: SIGPRO 6804

To appear in: *Signal Processing*

Received date: 16 November 2017
Revised date: 3 April 2018
Accepted date: 28 April 2018

Please cite this article as: Francisco J. Ibarrola, Leandro E. Di Persia, Ruben D. Spies, A Bayesian approach to convolutive nonnegative matrix factorization for blind speech dereverberation, *Signal Processing* (2018), doi: 10.1016/j.sigpro.2018.04.024

Highlights

- A dereverberation method that needs no pre-processing nor specific information.
- Very good performance observed in real-world recording conditions.
- Performance speed is fast enough to use as startpoint for on-line dereverberation.
- The theoretical ground is solid and allows for easily seeking ways of improvement.

A Bayesian approach to convolutive nonnegative matrix factorization for blind speech dereverberation

Francisco J. Ibarrola[a], Leandro E. Di Persia[a], Ruben D. Spies[b]

[a] *Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional, sinc(i), FICH-UNL/CONICET, Argentina. Ciudad Universitaria, CC 217, Ruta Nac. 168, km 472.4, (3000) Santa Fe, Argentina.*

[b] *Instituto de Matemática Aplicada del Litoral, IMAL, CONICET-UNL, Centro Científico Tecnológico CONICET Santa Fe, Colectora Ruta Nac. 168, km 472, Paraje "El Pozo", (3000), Santa Fe, Argentina and Departamento de Matemática, Facultad de Ingeniería Química, Universidad Nacional del Litoral, Santa Fe, Argentina.*

Abstract

When a signal is recorded in an enclosed room, it typically gets affected by reverberation. This degradation represents a problem when dealing with audio signals, particularly in the field of speech signal processing, such as automatic speech recognition. Although there are some approaches to deal with this issue that are quite satisfactory under certain conditions, constructing a method that works well in a general context still poses a significant challenge. In this article, we propose a Bayesian approach based on convolutive nonnegative matrix factorization that uses prior distributions in order to impose certain characteristics over the time-frequency components of the restored signal and the reverberant components. An algorithm for implementing the method is described and tested. Comparisons of the results against those obtained with state-of-the-art methods are presented, showing significant improvement.

Keywords: signal processing, dereverberation, regularization

1. Introduction

In recent years, many technological developments have attracted attention towards human-machine interaction. Since the most natural and easiest way of

Email address: fibarrola@sinc.unl.edu.ar (Francisco J. Ibarrola)

human communication is through speech, much research effort has been put into achieving the same natural interaction with machines. This effort has already generated many advances in a wide variety of fields such as automatic speech recognition ([1]), automatic translation systems ([2]) and control of remote devices through voice ([3]), to name only a few. A significant amount of work has been recently devoted to produce robustness in speech recognition ([4]), resulting in several advances in the areas of speech enhancement ([1], [5]), multiple sources separation ([6], [7]), and particularly in dereverberation techniques ([8]), which constitute the topic of this work.

When recorded in enclosed rooms, audio signals will most certainly be affected by reverberant components due to reflections of the sound waves in the walls, ceiling, floor or furniture. This can severely degrade the characteristics of the recorded signal ([9]), generating difficult problems for its processing, particularly when required for certain speech applications ([10]). The goal of any dereverberation technique is to remove or to attenuate the reverberant components in order to obtain a cleaner signal. The dereverberation problem is called "blind" when the available data consists only of the reverberant signal itself, and this is the problem we shall deal with in this work.

Depending on the problem, our observation might consist of a single or multi-channel signal, that is, we might have a signal recorded by one or more microphones. For the latter case, quite a few methods exist that work relatively well ([11], [12]).

For the single-channel case, we may distinguish between supervised and unsupervised approaches. The first kind refers to those that begin with a training stage that serves to learn some characteristics of the reververation conditions, while the second kind alludes to those methods that can be implemented directly over the reverberant signal. Some supervised methods ([13], [14], [15]) appear to perform somewhat better than unsupervised ones, but they pose the disadvantage of needing learning data corresponding to the specific room conditions, microphone and source locations, and a previous process that might take a significant amount of time.

In the context of unsupervised blind dereverberation, although some recently proposed methods ([12], [16]) work reasonably well, there is still much room for improvement. Our work is based on a convolutive non-negative matrix factorization (NMF) reverberation model, as proposed by Kameoka *et al* ([16]), along with a Bayesian approach for building a functional that takes into account *a priori* expected characteristics over the elements of the representation model. This functional can be thought of as the cost function of a mixed penalization model, such as in [17]. This kind of approach has been also recently used and successfully applied by several authors in many areas, mainly in signal and image processing applications ([18], [19], [20], [21], [22]). These techniques have shown to produce good results in terms of enhancing certain desirable characteristics on the solutions while precluding unwanted ones.

2. A Reverberation Model

Let $s, x : \mathbb{R} \to \mathbb{R}$, with support in $[0, \infty)$, be the functions associated to the clean and reverberant signals, respectively. As it is customary, we shall assume that the reverberation process is well represented by a Linear Time-Invariant (LTI) system. Thus, the reverberation model can be written as

$$x(t) = (h * s)(t), \tag{1}$$

where $h : \mathbb{R} \to \mathbb{R}$ is the room impulse response (RIR) signal, and "$*$" denotes convolution. This LTI hypothesis implies we are assuming the source and microphone positions to be static, and the energy of the signal to be low enough for the effect of the non-linear components to be relatively insignificant.

When dealing with sound signals (particularly speech signals), it is often convenient to work with the associated spectrograms rather than the signals themselves. Thus, we make use of the short time Fourier transform (STFT), defined as

$$\mathbf{x}_k(t) \doteq \int_{-\infty}^{\infty} x(u) w(u-t) e^{-2\pi i u k} du, \quad t, k \in \mathbb{R},$$

where $w : \mathbb{R} \to \mathbb{R}_0^+$ is a compactly supported, even function such that $\|w\|_1 = 1$. This function is called *window*.

In practice, we work with discretized versions of the signals involved ($x[\cdot]$, $h[\cdot]$, $s[\cdot]$, and $w[\cdot]$). With this in mind, we shall define the discrete STFT as

$$\mathbf{x}_k[n] \doteq \sum_{m=-\infty}^{\infty} x[m]w[m-n]e^{-2\pi imk}, \quad n, k \in \mathbb{N}.$$

Denoting the STFTs of s and h by $\mathbf{s}_k[n]$ and $\mathbf{h}_k[n]$, respectively, a discretized approximation of the STFT model associated to (1) is given by

$$\mathbf{x}_k[n] \approx \tilde{\mathbf{x}}_k[n] \doteq \sum_{\tau=0}^{N_h-1} \mathbf{s}_k[n-\tau]\mathbf{h}_k[\tau], \tag{2}$$

where $n = 1, \ldots, N$, is a discretized time variable that corresponds to window location, $k = 1, \ldots, K$, denotes the frequency subband and N_h is a parameter of the model associated to the expected maximum duration of the reverberation phenomenon. The model is built as in [23], being the approximation due to the use of band-to-band filters only. Later on, the values of n will be chosen in such a way that the union of the windows' supports contain the support of the observed signal, and the values of k in such a way that they cover the whole frequency spectrum, up to half the sampling frequency.

Now, let us write $\mathbf{h}_k[\tau] = |\mathbf{h}_k[\tau]|e^{j\phi_k[\tau]}$. It is well known ([24]) that the phase angles $\phi_k[\tau]$ are highly sensitive with respect to mild variations on the reverberation conditions. To overcome the problems derived from this, we shall proceed (see [16]) treating the $K \times N_h$ variables $\phi_k[\tau]$ as *i.i.d.* random variables with uniform distribution in $[-\pi, \pi)$. Denoting the complex conjugate by "*"

and the Kronecker delta by δ_{ij}, the expected value of $|\tilde{\mathbf{x}}_k[t]|^2$ is given by

$$\begin{aligned}
E|\tilde{\mathbf{x}}_k[n]|^2 &= E\sum_{\tau,\tau'} \mathbf{s}_k[n-\tau]\mathbf{s}_k^*[n-\tau']\mathbf{h}_k[\tau]\mathbf{h}_k^*[\tau'] \\
&= E\sum_{\tau,\tau'} \mathbf{s}_k[n-\tau]\mathbf{s}_k^*[n-\tau']\,|\mathbf{h}_k[\tau]|\,e^{j\phi_k[\tau]}\,|\mathbf{h}_k[\tau']|\,e^{-j\phi_k[\tau']} \\
&= \sum_{\tau,\tau'} \mathbf{s}_k[n-\tau]\mathbf{s}_k^*[n-\tau']\,|\mathbf{h}_k[\tau]|\,|\mathbf{h}_k[\tau']|\,Ee^{j(\phi_k[\tau]-\phi_k[\tau'])} \\
&= \sum_{\tau,\tau'} \mathbf{s}_k[n-\tau]\mathbf{s}_k^*[n-\tau']\,|\mathbf{h}_k[\tau]|\,|\mathbf{h}_k[\tau']|\,\delta_{\tau\tau'} \\
&= \sum_{\tau} |\mathbf{s}_k[n-\tau]|^2\,|\mathbf{h}_k[\tau]|^2.
\end{aligned}$$

Note that the $[-\pi,\pi)$ interval choice for $\phi_k[\tau]$ is arbitrary, since this result holds for any $2\pi-$length interval. Finally, let us define $S_k[n] \doteq |\mathbf{s}_k[n]|^2$, $H_k[n] \doteq |\mathbf{h}_k[n]|^2$ and $X_k[n] \doteq E|\tilde{\mathbf{x}}_k[n]|^2$. Then, our model reads

$$X_k[n] = \sum_{\tau} S_k[n-\tau]H_k[\tau], \tag{3}$$

and the square magnitude of the observed spectrogram components can be written as

$$Y_k[n] = X_k[n] + \epsilon_k[n], \tag{4}$$

where $\epsilon_k[n]$ denotes the representation error. As shown in [16], this model is equivalent to a convolutive NMF ([25]) with diagonal basis. In the next section, we derive a cost function in order to find an appropriate convolutive representation that allows us to isolate the components $S_k[n]$.

3. A Bayesian approach

In the following, we will use a Bayesian approach to derive a cost function which we will then minimize in order to obtain our regularized solution. Let us begin by assuming, for every k, $\epsilon_k[n], S_k[n], H_k[n]$ are independent random variables, also independent with respect to k. Also, let us denote by $S, Y, X \in \mathbb{R}^{K\times N}$ and $H \in \mathbb{R}^{K\times N_h}$ the non-negative matrices whose (k,n)-th elements are $S_k[n], Y_k[n], X_k[n]$ and $H_k[n]$, respectively.

More often than not, some type of "patterns" can be observed in a speech spectrogram, mainly due to the harmonics of speech (see Figure 1). However, they seem to be strongly speaker and phoneme dependent, and although it would be interesting to try to model this correlation, this is not viable in a blind setting (since no a-priori information is available for estimating it). Besides, it is worth mentioning that the frequency independency assumption has shown to lead to quite good results.

As it is customary ([16]), for the representation error, we assume $\epsilon_k[n] \sim \mathcal{N}(0, \sigma_k^2)$, where $\sigma_k > 0$ is an unknown parameter, and the variables are non-correlated with respect to n. Hence, it follows from (4) that the conditional distribution of Y given S and H (i.e. the likelihood) is given by

$$\pi_{like}(Y|S,H) = \prod_{k=1}^{K}\prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(Y_k[n]-X_k[n])^2}{\sigma_k^2}\right).$$

Note that, strictly speaking, in the above model for the representation error, the non-negativity constraint on the components of Y is not enforced. This is done mainly for simplicity reasons. It is rooted in the fact that this distribution provides a good model for the data Y; thus, the probability of one of its components be negative is very small, and enforcing non-negativity would unnecessarily complicate the model.

Let us now turn our attention to S. Figure 1 depicts the log-spectrograms for a clean signal and its reverberant version. As it can be observed, while the spectrogram of the clean signal is somewhat sparse, the one corresponding to the reverberant signal presents a smoother or more diffuse structure. The presence of discontinuities in the spectrogram of the clean signal can be favored by assuming S follows a generalized non-negative Gaussian distribution ([26]). Thus,

$$\pi_{prior}(S) = \begin{cases} \prod_{k=1}^{K}\prod_{n=1}^{N} \frac{1}{\Gamma(1+1/p)b_k} \exp\left(-\frac{S_k[n]^p}{b_k^p}\right) & S_k[n] \geq 0, \\ 0 & S_k[n] < 0, \end{cases}$$

where $p \in (0,2)$ is a prescribed parameter and $b_k > 0$ is unknown.

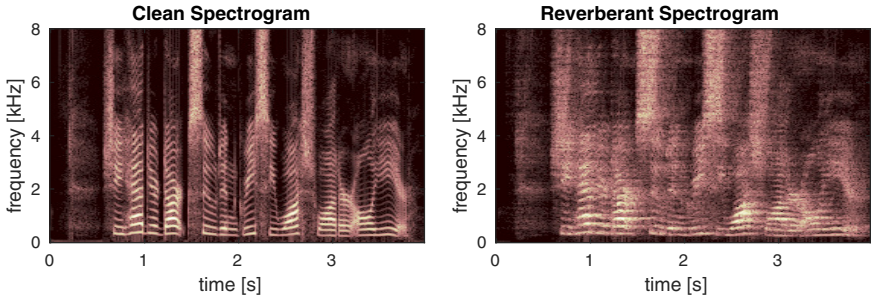

Figure 1: Spectrograms for a clean speech signal (left) and the corresponding reverberant speech signal (right). The clean signal, from the TIMIT database, was sampled at 16 [kHz], and corresponds to a female voice uttering the sentence 'She had your dark suite in greasy wash water all year.' The signal was artificially made reverberant by convolution with a room impulse response, with a reverberation time of 600 [ms], to produce the reverberant spectrogram. Both spectrograms were made using Hamming windows with 512 samples and an overlapping of 256.

In regards to H, although no general conditions are expected on its individual components, we do expect its first order time differences to exhibit a certain degree of regularity (see Figures 2 and 3). It can be observed that the log-spectrograms consist of a high-energy vertical band to the left, that corresponds to the linear impulse response, and some straight lines of less energy that correspond to the non-linear distortions produced by the increase on the rate at which the echoes reach the receiver ([27]). In fact, if windows are set close enough relative to the duration of the reverberation phenomenon, then consecutive time components of H will capture overlapped information, which along with the exponential decay characteristic of the RIR ([28]) accounts for a somewhat smooth structure. Therefore, we define the time differences matrix $V \in \mathbb{R}^{K\times(N_h-1)}$, with components $V_k[n] \doteq H_k[n] - H_k[n-1]\ \forall n = 1,\ldots,N_h-1,\ k = 1,\ldots,K$. The regularity of these variations is contemplated by assuming V follows a normal distribution with zero mean and variance η_k^2:

$$\pi_{prior}(V) = \prod_{k=1}^{K}\prod_{n=2}^{N_h} \frac{1}{\sqrt{2\pi}\eta_k} \exp\left(-\frac{V_k[n]^2}{\eta_k^2}\right).$$

Let $H_k \in \mathbb{R}^{N_h}$ be the transpose k^{th}-row of H, $L \in \mathbb{R}^{N_h-1\times N_h}$ be the matrix

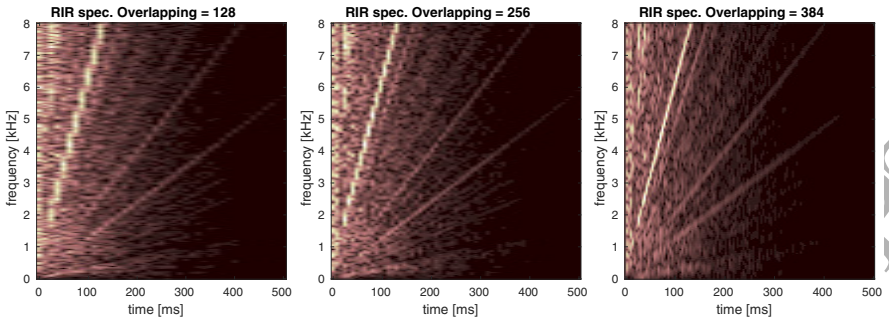

Figure 2: Log-spectrograms for an artificial 16 [kHz] RIR signal with reverberation time of 600 [ms]. The spectrograms were made using a hamming window length of 512 and different overlappings.

such that $LH_k = V_k$ and $\pi_{prior}(H)$ the prior induced from $\pi_{prior}(V)$ through this relation. Using Bayes' theorem, the *a posteriori* joint distribution of S and H conditioned to Y satisfies

$$\pi_{post}(S, H|Y) \propto \pi_{like}(Y|S, H)\pi_{prior}(S)\pi_{prior}(H). \tag{5}$$

Our goal is to find $\hat{S}$ and $\hat{H}$ that are representative of the *a posteriori* distribution (5). Although the immediate instinct might be to compute the expected value, there are quite a few other ways to proceed, with different degrees of reliability and complexity. In the light of the assumed distributions and the high dimensionality of the problem, the *maximum a posteriori* (MAP) estimator is a reasonable choice in this case. Note that maximizing (5) is tantamount to minimizing $-\log \pi_{post}(S, H|Y)$. If we denote by $S_k, Y_k, X_k \in \mathbb{R}^N$, $H_k \in \mathbb{R}^{N_h}$ and $V_k \in \mathbb{R}^{N_h - 1}$ the (transposed) rows of S, Y, X, H and V, then

$$\begin{aligned} J(S, H) &\doteq -\log \pi_{post}(S, H|Y) \\ &= \sum_{k=1}^{K} \left[\frac{1}{\sigma_k^2} ||Y_k - X_k||_2^2 + \frac{1}{b_k^p} \sum_n S_k[n]^p + \frac{1}{\eta_k^2} ||LH_k||_2^2 \right] + C, \end{aligned} \tag{6}$$

where C is a constant independent of S and H. Our goal is to minimize J, subject to the non-negativity restrictions $S_k[n] \geq 0 \, \forall k = 1, \ldots, K$, $n = 1, \ldots, N$, $H_k[n] \geq 0 \, \forall k = 1, \ldots, K$, $n = 1, \ldots, N_h$.

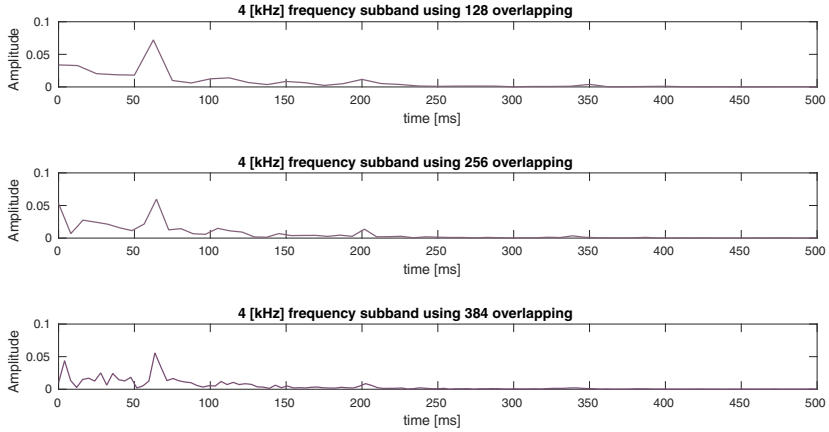

Figure 3: Signals corresponding to the 4[kHz] frequency subband of RIR spectrograms $H_{129}[n]$, $n = 1, \ldots, N$, with window length 512 and different overlappings. The sampling frequency is of 16[kHz] and the reverberation time is 600 [ms]. The signals show certain regularity, which increases with the window overlapping.

Although it is likely that different frequency sub-bands be affected differently by the RIR, with the reverberant spectrogram being the only available data for a blind approach, there will always be an arbitrary frequency dependent scaling ambiguity. In this way, it is impossible to exactly recover the original scaling of the source. Since given this fundamental indeterminacy, any frequency bin amplitude would be arbitrary in some sense, we have imposed the constraint $||S_k||_\infty = ||Y_k||_\infty \; \forall k$, which means that the maximum values shall remain equal for every frequency bin (this is similar to the minimum distortion principle ([29]) applied in frequency domain blind source separation). Additionally, we have experimentally found this constraint to be adequate.

3.1. Model parameters

Before proceeding to minimize equation (6), some comments on the model parameters $\{\sigma_k, b_k, \eta_k, p\}_{k=1,\ldots,K}$ are in order.

The value of the exponent $p \in (0, 2)$ is related to the degree of sparsity of S. While small values of p will promote high sparsity, choosing $p \approx 2$ will yield low sparsity.

Notice that for any given $k \in \{1, \ldots, K\}$, the variance of the representation error is proportional to the energy (the square of the L^2-norm) of the corresponding frequency sub-band. That is, we choose $\sigma_k^2 \doteq \sigma_0^2 \|Y_k\|^2$, where σ_0 is a constant independent of k. In a similar fashion, we choose $b_k \doteq b_0 \|Y_k\|$. Finally, since we have no evidence of any relationship between the frequency sub-band and the variations of H, we choose $\eta_k \doteq \eta_0$, independent of the frequency bin. Furthermore, since the functional (6) can be minimized separately in each frequency bin, the selection of the parameters is simplified by first choosing p and then the ratios σ_0^2/b_0^p and σ_0^2/η_0^2.

4. Hypermodel approach

To better deal with uncertainty on some of the parameter values, the previous model can be extended to a hypermodel by considering those parameters as random variables. For instance, due to the aforementioned uncertainty on the variance of H, we shall assume that the standard deviations of H_k, $\eta_k > 0, k = 1, \ldots, K$, are realizations of *i.i.d.* random variables with gamma distribution. That is,

$$\pi_{hyper}(\eta_k) \doteq \frac{\eta_k^{\alpha-1}}{\beta^\alpha \Gamma(\alpha)} \exp\left(-\frac{\eta_k}{\beta}\right),$$

where $\alpha > 1$ and $\beta > 0$ are shape and scale parameters, respectively. Using this hyperprior, the new functional (the negative logarithm of the *a-posteriori* distribution) turns out to be:

$$\begin{aligned} J_{hyp}(S, H, \eta) &\doteq -\log \pi_{post}(S, H, \eta | Y) \\ &= \sum_{k=1}^{K} \left[\frac{1}{\sigma_k^2} ||Y_k - X_k||_2^2 + \frac{1}{b_k^p} \sum_n S_k[n]^p + \frac{1}{\eta_k^2} ||LH_k||_2^2 \right] \\ &+ \sum_{k=1}^{K} \left[(N_h + 1 - \alpha) \log \eta_k + \frac{\eta_k}{\beta} \right] + C, \end{aligned} \tag{7}$$

where η denotes the vector whose components are $\eta_k, k = 1, \ldots, K$ and C is a constant independent of S, H, and η.

In what follows, we focus on minimizing the functionals J and J_{hyp} defined by (6) and (7), respectively.

5. Iterative minimization algorithms

5.1. Minimizing J

We begin by introducing a method for minimizing J, defined in (6). Later on, we will show that by adding an extra step, the same method can be used for minimizing J_{hyp}.

5.1.1. Auxiliary functions

The algorithm is constructed based on an auxiliary function technique, following similar ideas as those in [16]. Minimization procedures based in this kind of techniques are also known as Majorization-Minimization algorithms ([30]).

Let $\Omega \subset \mathbb{R}$ and $f : \Omega \to \mathbb{R}_0^+$. Then, $g : \Omega \times \Omega \to \mathbb{R}_0^+$ is called an *auxiliary function* for f if

$$(i)\ g(w,w) = f(w) \ \text{ and } \ (ii)\ g(w,w') \geq f(w), \ \ \forall w, w' \in \Omega. \tag{8}$$

Let $w^0 \in \Omega$ be arbitrary and let

$$w^j \doteq \arg\min_w g(w, w^{j-1}). \tag{9}$$

With this definition, it can be shown ([31]) that the sequence $\{f(w^j)\}_j$ is non-increasing. We intend to use this property as a tool for alternatively updating the matrices H and S. Let us begin by fixing $H = H'$, where H' is an arbitrary $K \times N_h$ matrix. Then, an auxiliary function for $J(S, H')$ (as defined in (6)) with respect to S is given by

$$\begin{aligned} g_s(S, S') \doteq & \sum_{k,n,\tau} \frac{1}{\sigma_k^2} \frac{S'_k[\tau] H'_k[n-\tau]}{X'_k[n]} \left(Y_k[n] - \frac{S_k[\tau]}{S'_k[\tau]} X'_k[n] \right)^2 + \sum_k \frac{1}{\eta_k^2} ||L H'_k||_2^2 \\ & + \sum_{k,n} \frac{1}{b_k^p} \left(\frac{p}{2} S'_k[n]^{p-2} S_k[n]^2 + S'_k[n]^p - \frac{p}{2} S'_k[n]^p \right), \end{aligned} \tag{10}$$

where $X'_k[n] = \sum_\tau S'_k[n-\tau] H'_k[\tau]$. The proof can be found in Appendix A.

In an analogous way, it can be shown that if we let $S = S'$ be fixed, where

S' is an arbitrary $K \times N$ matrix, then

$$g_h(H, H') \doteq \sum_{k,n,\tau} \frac{1}{\sigma_k^2} \frac{S'_k[n-\tau]H'_k[\tau]}{X'_k[n]} \left(Y_k[n] - \frac{H_k[\tau]}{H'_k[\tau]} X'_k[n] \right)^2 + \sum_k \frac{1}{b_k^p} ||S'_k||_p^p + \sum_k \frac{1}{\eta_k^2} ||LH_k||_2^2$$

is an auxiliary function for $J(S', H)$ with respect to H.

Having defined auxiliary functions, we will use the updating rule derived from (9) to build an algorithm for iteratively updating matrices S and H in order to minimize J. Notice this requires minimizing g_s and g_h with respect to the updating variables, but since g_s is quadratic with respect to S and g_h is quadratic with respect to H, we can simply use the first order necessary conditions in both cases. From this point on, in the context of the iterative updating process, S' and H' will refer not to arbitrary nonnegative matrices, but to those estimations of S and H obtained in the immediately previous step.

5.1.2. Updating rule for S

Firstly, we shall derive an updating rule for $S_k[\tau]$. That is, we wish to minimize g_s w.r.t. S. The first order necessary condition on g_s yields

$$\begin{aligned}
0 =& \frac{\partial g_s(S, S')}{\partial S_k[\tau]} \\
=& -2 \sum_n \frac{1}{\sigma_k^2} H'_k[n-\tau] \left(Y_k[n] - \frac{S_k[\tau]}{S'_k[\tau]} X'_k[n] \right) + \frac{p}{b_k^p} S'_k[\tau]^{p-2} S_k[\tau] \\
=& -\sum_n H'_k[n-\tau] Y_k[n] + \frac{S_k[\tau]}{S'_k[\tau]} \sum_n H'_k[n-\tau] X'_k[n] + \frac{p\sigma_k^2}{2b_k^p} S'_k[\tau]^{p-2} S_k[\tau] \\
=& -S'_k[\tau] \sum_n H'_k[n-\tau] Y_k[n] + \left(\sum_n H'_k[n-\tau] X'_k[n] + \frac{p\sigma_k^2}{2b_k^p} S'_k[\tau]^{p-1} \right) S_k[\tau],
\end{aligned}$$

which easily leads to the multiplicative updating rule

$$S_k[\tau] = S'_k[\tau] \frac{\sum_n H'_k[n-\tau] Y_k[n]}{\sum_n H'_k[n-\tau] X'_k[n] + \frac{p\sigma_k^2}{2b_k^p} S'_k[\tau]^{p-1}}.$$

In order to avoid the aforementioned scale indeterminacy, every updating step is to be followed by scaling S_k so that its ℓ^∞ norm coincides with that of the observation Y_k.

5.1.3. Updating rule for H

In order to state an updating rule for H, we begin by defining the diagonal matrices $A^k, B^k \in \mathbb{R}^{N_h \times N_h}$, whose diagonal elements are $A^k_{\tau,\tau} \doteq \sum_n S'_k[n-\tau]X'_k[n]$ and $B^k_{\tau,\tau} \doteq H'_k[\tau]$, and the vector $\zeta^k \in \mathbb{R}^{N_h}$ with components $\zeta^k_\tau = \sum_n S'_k[n-\tau]Y_k[n]$.

It can be shown (see Appendix B) that with these definitions, H can be updated by solving the linear system

$$\left(A^k + \frac{\sigma_k^2}{\eta_k^2} B^k L^{\mathrm{T}} L\right) H_k = B^k \zeta^k. \tag{11}$$

Let us notice that under the assumption that the diagonal elements of A^k and B^k are strictly positive, and since $L^{\mathrm{T}}L$ is positive-semidefinite, $(B^k)^{-1}A^k + \lambda_{h,k} L^{\mathrm{T}}L$ is positive-definite, and hence the linear system has a unique solution. Furthermore, this implies that the solution is non-negative. The assumption of $A^k_{\tau,\tau} > 0$ is adequate, since these elements correspond to the discrete convolution of S'_k and X'_k. Although the validity of the hypothesis over $B^k_{\tau,\tau}$ is not so clear, in practice, the matrix in system (11) has turned out to be non-singular. Nonetheless, H_k can be computed as the best approximate solution in the least-squares sense. Solving this $N_h \times N_h$ linear system entails no challenge, since N_h is usually chosen relatively small, depending on the window step and the reverberation time.

5.2. Minimizing J_{hyp}

It follows immediately from the fact that the additional terms on equation (7) with respect to equation (6) do not depend on S nor H, that the minimization steps derived for J are suitable for J_{hyp} as well. Thus, it only remains to minimize J_{hyp} with respect to η, which can be shown (see Appendix C) to be equivalent to solving the following equation:

$$\eta_k^3 + (N_h + 1 - \alpha)\beta\, \eta_k^2 - 2\beta ||LH_k||_2^2 = 0,$$

for every $k = 1, \ldots, K$. This can be done either explicitly by means of the general solution of the cubic equation, or by an appropriate iterative method.

5.3. Final considerations

All steps of the dereverberation process are stated in Algorithm 1. The updating step in line 22 only concerns functional J_{hyp}, and it must be skipped when minimizing J.

In the Initialization Step we define the clean spectrogram S equal to the observation, which is natural since in a way they both correspond to the same signal, and H_k as a vector with exponential time decay, which is an expected characteristic of a RIR. Note that with this initialization all the variables result non-negative. Under this condition, it is easy to see that all the updating rules maintain non-negativitiy, thus complying with the aforementioned restrictions $S_k[n] \geq 0 \, \forall k = 1, \ldots, K$, $n = 1, \ldots, N$, and $H_k[n] \geq 0 \, \forall k = 1, \ldots, K$, $n = 1, \ldots, N_h$..

Finally, we set the stopping criterion over the decay of the norm of two consecutive approximations of S. This has shown to work quite well, although other stopping criteria might be considered.

Results to illustrate the performance of the proposed algorithms are presented in the next section.

6. Experimental results

For the experimental results we used both simulated and recorded reverberant signals. While a large number of artificially reverberant signals were produced to get statistically significant results, recorded signals were used to corroborate the performance of the methods using real data.

6.1. Experiments with simulations

For the experiments, we took 110 speech signals from the TIMIT database ([32]), recorded at 16 kHz, and artificially made them reverberant by convolution with impulse responses generated with the software Room Impulse Response Generator[1], based on the model in [33]. Each signal was degraded under differ-

[1] https://github.com/ehabets/RIR-Generator

Algorithm 1 Bayesian dereverberation

Initializing

$S \leftarrow Y$

$H_k[n] \leftarrow \exp(-n) \quad \forall k = 1 \dots K,\ n = 1 \dots N$

MAIN LOOP

for $i = 1 \dots \text{maxiter}$

  $X_k[n] \leftarrow \sum_\tau S_k[n-\tau] H_k[\tau] \quad \forall k = 1 \dots K,\ n = 1 \dots N$

  for $k = 1 \dots K$

    for $\tau = 1 \dots N$

      $S_k[\tau] \leftarrow S_k[\tau] \dfrac{\sum_n H_k[n-\tau] Y_k[n]}{\sum_n H_k[n-\tau] X_k[n] \ + \ \frac{p\sigma_k^2}{2b_k^p} S_k[\tau]^{p-1}}.$

    end for

    $S_k \leftarrow S_k \dfrac{\|Y_k\|_\infty}{\|S_k\|_\infty}.$

  end for

  for $k = 1 \dots K$

    Build the diagonal matrices $A^k, B^k \in \mathbb{R}^{N_h \times N_h}$:

      $A^k_{\tau,\tau} = \sum_n S_k[n-\tau] X_k[n],$

      $B^k_{\tau,\tau} = H_k[\tau].$

    Build the vector ζ^k :

      $\zeta^k_\tau = \sum_n S_k[n-\tau] Y_k[n]$

    Solve for H_k :

      $(A^k + \dfrac{\sigma_k^2}{\eta_k^2} B^k L^{\mathrm{T}} L) H_k = B^k \zeta^k.$

    if Using the hypermodel (J_{hyp})

      Solve for η_k : $\quad \eta_k^3 + (N_h + 1 - \alpha)\beta\, \eta_k^2 - 2\beta ||LH_k||_2^2 = 0.$

    end if

  end for

  if $\|S - S'\|_F \leq \delta$

    return

  end if

end for

ent reverberation conditions: three different room sizes, each with three different microphone positions and four different reverberation times, which gives us a total of 3960 signals for testing. Table 1 gives account of the room dimensions and source and microphone positions that were chosen.[2]

Table 1: Simulated room settings

	Length	Width	Height
Room 1 dimensions	5.00 [m]	4.00 [m]	6.00 [m]
Room 2 dimensions	4.00 [m]	4.00 [m]	3.00 [m]
Room 3 dimensions	10.0 [m]	4.00 [m]	5.00 [m]
Source position	2.00 [m]	3.50 [m]	2.00 [m]
Microphone 1 position	2.00 [m]	1.50 [m]	1.00 [m]
Microphone 2 position	2.00 [m]	2.00 [m]	1.00 [m]
Microphone 3 position	2.00 [m]	2.00 [m]	2.00 [m]

In order to avoid preprocessing, the choice of the probabilistic model parameters was made *a priori* by means of empirical rules, based upon signals from a different database. This is supported by the fact that the parameters were observed to be rather robust with respect to variations of the reverberation conditions, and hence they were chosen simply as $\sigma_k^2 = \|Y_k\|^2$, $\eta_k = 1$ and $b_k = \|Y_k\| \times 10^7$. For the case of minimizing functional J_{hyp}, we set $\alpha = 10^2$ and $\beta = 10^{-2}$, so the expected value for η_k is $\alpha\beta = 1$, for the comparison between the Bayesian model and Hypermodel to be fair. The rest of the model parameters were chosen as specified in Table 2.

Table 2: Model parameter values

p	N_h	win.	window size	win. overlap.	δ	max. iter.
1	15	Ham.	512 samples	256 samples	$\|Y\|_F \times 10^{-3}$	20

Let us point out that the choice of N_h was done as to allow H to capture

[2]A web demo can be found in sinc.unl.edu.ar/web-demo/blindder/

early reverberation while precluding overlapped representations. In the first place, it is desirable for H to represent the RIR along the full Early Decay Time (EDT), the time period in which the reverberation phenomenon alters the clean signal the most, so its effect can be nullified. On the other hand, if we were to choose N_h too large, it might lead certain similarities in the observation Y within a fixed frequency range to be represented as echoes from high energy components of S. It is worth mentioning, however, that the performance of our dereverberation method has shown no high sensitivity with respect to the choice of N_h.

In order to evaluate the performance of our models, using both functionals J and J_{hyp}, we made comparisons against three state-of-the-art methods that work under the same conditions. Two of the methods we used were those proposed by Kameoka *et al* in [16] and the mixed penalization method proposed in [17], which are not only recent but in a sense precursors to the method proposed in this article. Also, we included the method proposed by Wisdom *et al* in [12], with a window length of 2048, because of its great performance in the Reverb Challenge ([34]).

To measure performance, following [35], we made use of the frequency weighted segmental signal-to-noise ratio (fwsSNR) and cepstral distance. Furthermore, we also measured the speech-to-reverberation modulation energy ratio (SRMR, [36]), which has the advantage of being non-intrusive (it does not use the clean signal as an input). The results for each performance measure are stated in Table 3, and depicted in Figures 4- 6, classified in function of the reverberation times: 300[ms], 450[ms], 600[ms] and 750[ms]. Notice that for the cases of fwsSNR and SRMR, higher values correspond to better performance, while for the cepstral distance, small values indicate higher quality.

Table 3 shows that the results obtained using the Bayesian methods with functionals J and J_{hyp} are significantly better ($p < 0.01$) than those produced by the other methods for all the considered performance measures. Also, Figures 4-6 clearly show that in all cases the improvement is more evident for larger reverberation times, specially for the fwsSNR and the Cepstral Distance.

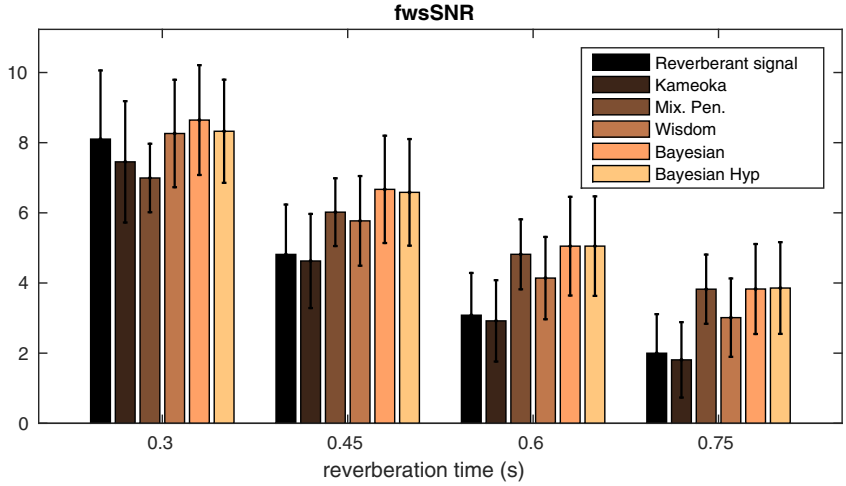

Figure 4: Mean and standard deviations of fwsSNR for different reverberation times.

Furthermore, Figure 5 shows that no competing method is able to reduce the Cepstral Distance for a reverberation time of 300[ms]. This most likely occurs because the reverberation time is too short and therefore the introduced distortion, when doing dereverberation, cancels out the potential gains. Yet, for larger reverberation times, our method does produce a significant improvement as measured by the Cepstral Distance. It is timely to mention that all the differences between the performance of our methods and every competing one hold statistical significance ($p < 0.01$) for every reverberation time (as depicted in Figures 4-6), with the only exception of the SRMR with a 300[ms] reverberation time, where our methods produce no significant improvement with respect to Wisdom's.

6.2. Experiments with recorded signals

For this experiment we have used real recordings obtained in our own office rooms, with a sampling frequency of 16[kHz]. Two male and two female speakers were randomly selected from the TIMIT database, and 10 speech signals for each

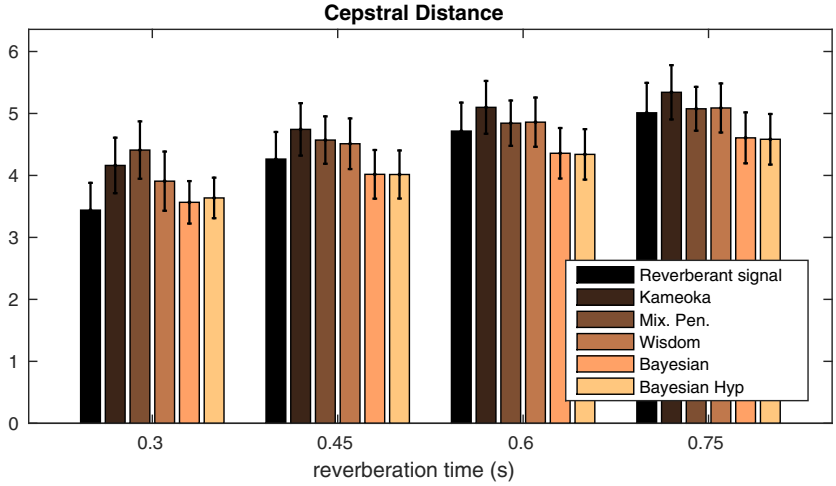

Figure 5: Mean and standard deviations of Cepstral Distance for different reverberation times.

were played in two different rooms. The dimensions of the fully furnished rooms and microphone positions are specified in Table 4. The reverberation times, measured using sine sweeps ([37]), were found to be 460[ms] on the first room and 440[ms] on the second. It is timely to mention that for the recordings to be realistic, they were made during standard office hours, with people working in nearby offices (although no people were present in the recording room), and some of the computers and air conditioning were left on.

The model parameters were chosen equal to those used for the experiment with simulations, except for the variance of the distribution of S, that was changed to cope with the considerably high noise level. The new choice was simply $b_k = 10\|S_k\|/\sigma_n$, where σ_n is the standard deviation of the noise, estimated from the first 1000 samples (61[ms]) of the recordings. The parameters for the competing methods were properly adjusted to the noise level as well.

Results are depicted in Table 5. Once again, we see that the Bayesian methods outperform the other methods in terms of the fwsSNR and SRMR, although Wisdom's method performs slightly better (but not significantly, $p >$

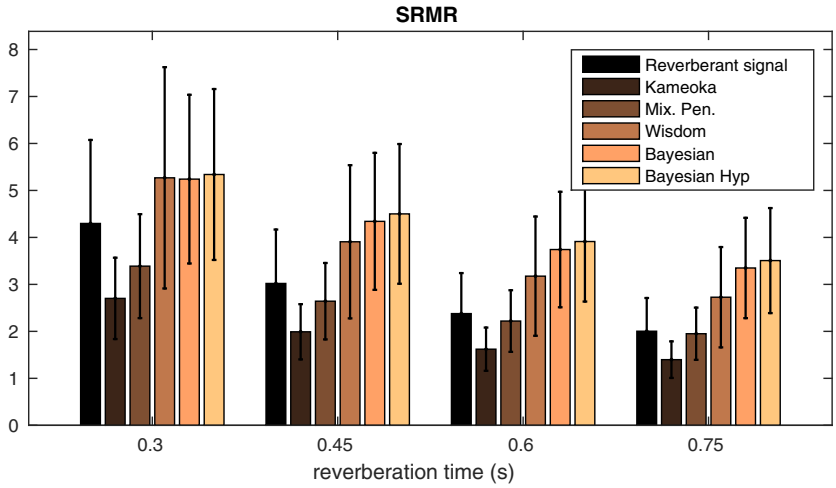

Figure 6: Mean and standard deviations of SRMR for different reverberation times.

0.01) in terms of Cepstral Distance.

6.3. Computing performance

Finally, we also compared the computing performance of the aforementioned methods using the TIMIT database of the first experiment. The examples were run using MatLab in a PC with an Intel Core i7-2600k CPU @3.4GHz×8, with 8Gb of RAM. The CPU-times for each method are depicted in Table 6, where it can be seen that although not as fast as the Mixed Penalization method, it is twice as fast as the closest competing method in terms of restoration quality. Finally, it is appropriate to mention that the speed of our method could be further improved using parallel computing. This is due to the fact that in our algorithm (just as in Kameoka's) the minimization can be performed simultaneously in every frequency bin.

7. Conclusions

In this work a new blind dereverberation method for speech signals based on a Bayesian approach over a convolutive NMF representation of the spectrograms

Table 3: Mean and standard deviation (between parenthesis) of performance measures for each method, using simulations. Best results are shown in boldface.

Measure	fwsSNR	Cepstral Dist.	SRMR
Reverberant	4.499 (2.73)	4.358 (0.75)	2.924 (1.48)
Kameoka	4.203 (2.52)	4.836 (0.62)	1.928 (0.78)
Mixed Pen	5.414 (1.55)	4.723 (0.47)	2.550 (0.98)
Wisdom	5.296 (2.35)	4.592 (0.61)	3.770 (1.91)
Bayesian	**6.048** (2.32)	**4.137** (0.55)	4.168 (1.58)
Hypermodel	5.954 (2.20)	4.144 (0.52)	**4.315** (1.60)

Table 4: Office rooms settings

	Length	Width	Height
Room 1 dimensions	4.15 [m]	3.00 [m]	3.00 [m]
Source 1 position	3.60 [m]	1.50 [m]	1.50 [m]
Microphone 1 position	1.10 [m]	1.50 [m]	1.50 [m]
Room 2 dimensions	5.85 [m]	4.55 [m]	3.00 [m]
Source 2 position	1.10 [m]	1.50 [m]	1.50 [m]
Microphone 2 position	1.10 [m]	4.00 [m]	1.50 [m]

was introduced and tested. This includes a basic Bayesian model as well as a model with hyperpriors.

Results show the new introduced method is faster and outperforms the others in terms of fwsSNR and SRMR, and, moreover, it is comparable to the best of those in terms of Cepstral Distance. A significant improvement in performance stands out for high reverberation times.

It is also worth mentioning that the proposed algorithm results fast enough to be considered for performing on-line dereverberation, endeavor that we plan to engage on in future work.

There is certainly much room for further improvement. Among others, the use of other prior distributions depending on *a-priori* information, the introduction of time variability, and exploring the use of other time-frequency represen-

Table 5: Mean and standard deviation (between parenthesis) of performance measures for each method. Best results are shown in boldface.

Measure	fwsSNR	Cepstral Dist.	SRMR
Reverberant	5.411 (3.23)	5.521 (0.87)	2.755 (0.75)
Kameoka	6.041 (3.19)	5.125 (0.68)	2.126 (0.48)
Mixed Pen	7.089 (3.19)	5.735 (0.79)	2.45 (0.58)
Wisdom	6.241 (3.60)	**4.640** (0.51)	3.227 (0.77)
Bayesian	8.608 (2.83)	4.839 (0.47)	4.860 (1.13)
Hypermodel	**8.660** (2.92)	4.824 (0.41)	**4.878** (1.14)

Table 6: Mean CPU time for dereverberation with each algorithm.

Method	Kameoka	Mixed Pen	Wisdom	Bayesian	Hyper.
CPU time	7.61[s]	4.15 [s]	11.14[s]	5.47[s]	5.58[s]

tations analogous to STFT that could help to improve the obtained restorations.

Acknowledgements

This work was supported in part by Consejo Nacional de Investigaciones Científicas y Técnicas, CONICET through PIP 2014-2016 Nº 11220130100216-CO, the Air Force Office of Scientific Research, AFOSR/SOARD, through Grant FA9550-14-1-0130, by Universidad Nacional del Litoral, UNL, through CAID-UNL 2011 Nº 50120110100519 "Procesamiento de Señales Biomédicas" and CAI+D-UNL 2016, PIC 50420150100036LI "Problemas Inversos y Aplicaciones a Procesamiento de Señales e Imágenes".

Appendix A. Proof of the fact that g_s is an auxiliary function for J

We want to prove that g_s, defined as in (10), is an auxiliary function for J, defined in (6). That is, we must show that g_s complies with both conditions stated in (8) .

The equality condition (i) is rather straightforward. In fact,

$$\begin{aligned}
g_s(S,S) =& \sum_{k,n,\tau} \frac{1}{\sigma_k^2} \frac{S_k[\tau]H_k'[n-\tau]}{\sum_\nu S_k[\nu]H_k'[n-\nu]} \left(Y_k[n] - \frac{S_k[\tau]}{S_k[\tau]} \sum_\nu S_k[\nu]H_k'[n-\nu] \right)^2 \\
&+ \sum_k \frac{1}{\eta_k^2} ||LH_k'||_2^2 + \sum_{k,n} \frac{1}{b_k^p} \left(\frac{p}{2} S_k[n]^{p-2} S_k[n]^2 + S_k[n]^p - \frac{p}{2} S_k[n]^p \right) \\
=& \sum_{k,n,\tau} \frac{1}{\sigma_k^2} \frac{S_k[\tau]H_k'[n-\tau]}{\sum_\nu S_k[\nu]H_k'[n-\nu]} \left(Y_k[n] - \sum_\nu S_k[\nu]H_k'[n-\nu] \right)^2 \\
&+ \sum_k \frac{1}{\eta_k^2} ||LH_k'||_2^2 + \sum_{k,n} \frac{1}{b_k^p} S_k[n]^p \\
=& \sum_{k,n} \frac{1}{\sigma_k^2} \left(Y_k[n] - \sum_\nu S_k[\nu]H_k'[n-\nu] \right)^2 + \sum_k \frac{1}{\eta_k^2} ||LH_k'||_2^2 + \sum_{k,n} \frac{1}{b_k^p} S_k[n]^p \\
=& J(S,H').
\end{aligned}$$

To prove condition (ii) in (8) we begin by defining

$$\begin{aligned}
P_{k,n} \doteq& \sum_\tau \frac{S_k'[\tau]H_k'[n-\tau]}{X_k'[n]} \left(Y_k[n] - \frac{S_k[\tau]}{S_k'[\tau]} X_k'[n] \right)^2, \\
R_{k,n} \doteq& (Y_k[n] - \sum_\tau S_k[\tau]H_k'[n-\tau])^2,
\end{aligned}$$

and $Q : \mathbb{R}^+ \to \mathbb{R}$ such that $Q(x) \doteq \frac{p}{2} x^{p-2} S_k[n]^2 + x^p - \frac{p}{2} x^p$. With these definitions, we can write

$$g_s(S,S') = \sum_k \left(\sum_n \left(\frac{1}{\sigma_k^2} P_{k,n} + \frac{1}{b_k^p} Q(S_k'[n]) \right) + \frac{1}{\eta_k^2} ||LH_k'||_2^2 \right),$$

and

$$J(S,H') = \sum_k \left(\sum_n \left(\frac{1}{\sigma_k^2} R_{k,n} + \frac{1}{b_k^p} S_k[n]^p \right) + \frac{1}{\eta_k^2} ||LH_k'||_2^2 \right).$$

Hence, to prove that $g_s(S,S') \geq J(S,H')\ \forall S,S'$ it is sufficient to show that

$P_{k,n} \geq R_{k,n}$ and $Q(S'_k[n]) \geq S_k[n]^p \; \forall n = 1, \ldots, N, k = 1, \ldots, K$. In fact,

$$\begin{aligned}
P_{k,n} - R_{k,n} =& \sum_{\tau} \frac{S'_k[\tau]H'_k[n-\tau]}{X'_k[n]} \left(Y_k[n] - \frac{S_k[\tau]}{S'_k[\tau]} X'_k[n] \right)^2 \\
& - \left(Y_k[n] - \sum_{\tau} S_k[\tau]H'_k[n-\tau] \right)^2 \\
=& \sum_{\tau} \frac{H'_k[n-\tau]S_k[\tau]^2 X'_k[n]}{S'_k[\tau]} - \left(\sum_{\tau} S_k[\tau]H'_k[n-\tau] \right)^2 \\
=& \sum_{\tau,\nu} \frac{H'_k[n-\tau]S_k[\tau]^2 H'_k[n-\nu]S'_k[\nu]}{S'_k[\tau]} - \sum_{\tau,\nu} S_k[\tau]H'_k[n-\tau]S_k[\nu]H'_k[n-\nu] \\
=& \sum_{\tau,\nu} \left(\frac{H'_k[n-\tau]S_k[\tau]^2 H'_k[n-\nu]S'_k[\nu]}{S'_k[\tau]} - S_k[\tau]H'_k[n-\tau]S_k[\nu]H'_k[n-\nu] \right) \\
=& \sum_{\tau \neq \nu} \left(\frac{H'_k[n-\tau]S_k[\tau]^2 H'_k[n-\nu]S'_k[\nu]}{S'_k[\tau]} - S_k[\tau]H'_k[n-\tau]S_k[\nu]H'_k[n-\nu] \right) \\
=& \sum_{\tau<\nu} H'_k[n-\tau]H'_k[n-\nu] \left(\frac{S_k[\tau]^2 S'_k[\nu]}{S'_k[\tau]} - 2S_k[\tau]S_k[\nu] + \frac{S_k[\nu]^2 S'_k[\tau]}{S'_k[\nu]} \right) \\
=& \sum_{\tau<\nu} \frac{H'_k[n-\tau]H'_k[n-\nu]}{S'_k[\nu]S'_k[\tau]} \left(S_k[\tau]S'_k[\nu] - S_k[\nu]S'_k[\tau] \right)^2 \geq 0.
\end{aligned}$$

To prove that $Q(S'_k[n]) \geq S_k[n]^p$, we begin by noting that $Q \in \mathcal{C}^\infty(\mathbb{R}^+)$. Then, the first order necessary condition for Q yields

$$0 = \frac{\partial Q}{\partial x} = \frac{p(p-2)}{2} x^{p-3} S_k[n]^2 + px^{p-1} - \frac{p^2}{2} x^{p-1} = \frac{p(p-2)}{2} x^{p-1}(x^{-2} S_k[n]^2 - 1),$$

meaning the only point at which the derivative of Q equals zero is at $x = S_k[n]$. Furthermore, $\frac{\partial^2}{\partial x^2} Q(S_k[n]) = S_k[n]^{p-2}(2p - p^2) > 0 \; \forall p \in (0,2)$, meaning that $Q(S_k[n]) = S_k[n]^p$ is the global minimum of Q. This yields

$$\begin{aligned}
g_s(S, S') &= \sum_k \left(\sum_n \left(\frac{1}{\sigma_k^2} P_{k,n} + \frac{1}{b_k^p} Q(S'_k[n]) \right) + \frac{1}{\eta_k^2} ||LH'_k||_2^2 \right) \\
&\geq \sum_k \left(\sum_n \left(\frac{1}{\sigma_k^2} R_{k,n} + \frac{1}{b_k^p} S_k[n]^p \right) + \frac{1}{\eta_k^2} ||LH'_k||_2^2 \right) = J(S, H').
\end{aligned}$$

■

Appendix B. Derivation of updating rule for H

In order to derive the updating rule for H, we shall write g_h as a function of the transposed rows H_k. We begin by noting

$$\begin{aligned} g_h(H,H') &= \sum_{k,n,\tau} \frac{1}{\sigma_k^2} \frac{S_k'[n-\tau]H_k'[\tau]}{X_k'[n]} \left(Y_k[n] - \frac{H_k[\tau]}{H_k'[\tau]} X_k'[n] \right)^2 \\ &\quad + \sum_k \frac{1}{b_k^p} ||S_k'||_p^p + \sum_k \frac{1}{\eta_k^2} ||LH_k||_2^2 \\ &= \sum_{k,n,\tau} \frac{1}{\sigma_k^2} \frac{S_k'[n-\tau]H_k'[\tau]Y_k^2[n]}{X_k'[n]} - 2 \sum_{k,n,\tau} \frac{1}{\sigma_k^2} S_k'[n-\tau]Y_k[n]H_k[\tau] \\ &\quad + \sum_{k,n,\tau} \frac{1}{\sigma_k^2} \frac{S_k'[n-\tau]X_k'[n]H_k^2[\tau]}{H_k'[\tau]} \\ &\quad + \sum_k \frac{1}{b_k^p} ||S_k'||_p^p + \sum_k \frac{1}{\eta_k^2} ||LH_k||_2^2. \end{aligned}$$

Next, we recall the definition of the diagonal matrices $A^k, B^k \in \mathbb{R}^{N_h \times N_h}$, whose diagonal elements are $A^k_{\tau,\tau} \doteq \sum_n S_k'[n-\tau]X_k'[n]$ and $B^k_{\tau,\tau} \doteq H_k'[\tau]$, and the vector $\zeta^k \in \mathbb{R}^{N_h}$ with components $\zeta^k_\tau = \sum_n S_k'[n-\tau]Y_k[n]$. With these definitions, we can write

$$\begin{aligned} g_h(H,H') &= \sum_{k,n,\tau} \frac{1}{\sigma_k^2} \frac{S_k'[n-\tau]H_k'[\tau]Y_k^2[t]}{X_k'[n]} - 2 \sum_k \frac{1}{\sigma_k^2} H_k^{\mathrm{T}} \zeta^k \\ &\quad + \sum_k \frac{1}{\sigma_k^2} H_k^{\mathrm{T}} A^k (B^k)^{-1} H_k + \sum_k \frac{1}{b_k^p} ||S_k'||_p^p + \sum_k \frac{1}{\eta_k^2} H_k^{\mathrm{T}} L^{\mathrm{T}} L H_k. \end{aligned}$$

Now, the first order necessary condition for g_h with respect to H_k is given by

$$0 = \frac{\partial g_h(H,H')}{\partial H_k} = -\frac{2}{\sigma_k^2}\zeta^k + \frac{2}{\sigma_k^2} A^k (B^k)^{-1} H_k + \frac{2}{\eta_k^2} L^{\mathrm{T}} L H_k,$$

which readily leads to the linear system

$$\left(A^k + \frac{\sigma_k^2}{\eta_k^2} B^k L^{\mathrm{T}} L \right) H_k = B^k \zeta^k.$$

Appendix C. Updating rule for η

In order to derive the updating rule for $\eta_k, k = 1, \ldots, K$, we begin by noting that $-\log \pi_{post}(S,H,\eta|Y) \in \mathcal{C}^1(0,\infty)$ with respect to η_k, and hence a local min-

imum must corresponds to a point with derivative equal to zero. Differentiating (7) with respect to η_k, we obtain

$$\frac{\partial}{\partial \eta_k} - \log \pi_{post}(S, H, \eta | Y) = -\frac{2}{\eta_k^3} ||LH_k||_2^2 + \frac{N_h + 1 - \alpha}{\eta_k} + \frac{1}{\beta}.$$

The first order necessary condition over (7) is thus tantamount to

$$\eta_k^3 + (N_h + 1 - \alpha)\beta\, \eta_k^2 - 2\beta ||LH_k||_2^2 = 0.$$

By Descartes' rule, this polynomial has exactly one positive root η_0. Since $\lim_{\eta_k \to \infty} (-\log \pi_{post}(S, H, \eta | Y)) = \infty$ and $\lim_{\eta_k \to 0^+} (-\log \pi_{post}(S, H, \eta | Y)) = \infty$, then η_0 is the global minimizer.

References

[1] M. Kim, H.-M. Park, Efficient online target speech extraction using DOA-constrained independent component analysis of stereo data for robust speech recognition, Signal Processing 117 (2015) 126–137.

[2] S. Yun, Y. J. Lee, S. H. Kim, Multilingual speech-to-speech translation system for mobile consumer devices, IEEE Transactions on Consumer Electronics 60 (3) (2014) 508–516.

[3] R. Neßelrath, M. M. Moniri, M. Feld, Combining speech, gaze, and micro-gestures for the multimodal control of in-car functions, in: 12th IEEE International Conference on Intelligent Environments (IE), 2016, pp. 190–193.

[4] L. Di Persia, D. Milone, H. L. Rufiner, M. Yanagida, Perceptual evaluation of blind source separation for robust speech recognition, Signal Processing 88 (10) (2008) 2578–2583.

[5] C. E. Martnez, J. Goddard, L. E. Di Persia, D. H. Milone, H. L. Rufiner, Denoising sound signals in a bioinspired non-negative spectro-temporal domain, Digital Signal Processing 38 (2015) 22–31.

[6] L. Di Persia, D. Milone, M. Yanagida, Indeterminacy free frequency-domain blind separation of reverberant audio sources., IEEE Transactions on Audio, Speech and Language Processing 17 (2) (2009) 299–311.

[7] L. E. Di Persia, D. H. Milone, Using multiple frequency bins for stabilization of FD-ICA algorithms, Signal Processing 119 (2016) 162–168.

[8] A. Tsilfidis, J. Mourjopoulos, Signal-dependent constraints for perceptually motivated suppression of late reverberation, Signal Processing 90 (3) (2010) 959–965.

[9] I. J. Tashev, Sound capture and processing: practical approaches, John Wiley & Sons, 2009.

[10] X. Huang, A. Acero, H.-W. Hon, R. Reddy, Spoken language processing: A guide to theory, algorithm, and system development, Vol. 95, Prentice hall PTR Upper Saddle River, 2001.

[11] M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, N. Ito, K. Kinoshita, M. Espi, T. Hori, T. Nakatani, et al., Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the reverb challenge, in: Proceedings of REVERB Challenge Workshop, 2014.

[12] S. Wisdom, T. Powers, L. Atlas, J. Pitton, Enhancement of reverberant and noisy speech by extending its coherence, in: Proceedings of REVERB Challenge Workshop, 2014.

[13] M. Moshirynia, F. Razzazi, A. Haghbin, A speech dereverberation method using adaptive sparse dictionary learning, in: Proceedings of REVERB Challenge Workshop, 2014.

[14] X. Xiao, S. Zhao, D. H. H. Nguyen, X. Zhong, D. L. Jones, E.-S. Chng, H. Li, The NTU-ADSC systems for reverberation challenge 2014, in: Proceedings of REVERB Challenge Workshop, 2014.

[15] K. Nathwani, R. M. Hegde, Joint source separation and dereverberation using constrained spectral divergence optimization, Signal Processing 106 (2015) 266–281.

[16] H. Kameoka, T. Nakatani, T. Yoshioka, Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms, in: 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, 2009, pp. 45–48.

[17] F. Ibarrola, L. Di Persia, R. Spies, Blind speech dereverberation using convolutive nonnegative matrix factorization with mixed penalization., in: Proceedings of VI Congreso de Matemática Aplicada, Computacional e Industrial, 2017, pp. 404–407.

[18] F. Ibarrola, G. Mazzieri, R. Spies, K. Temperini, Anisotropic BV-L^2 regularization of linear inverse ill-posed problems, Journal of Mathematical Analysis and Applications 450 (1) (2017) 427–443.

[19] D. Lazzaro, L. B. Montefusco, S. Papi, Blind cluster structured sparse signal recovery: A nonconvex approach, Signal Processing 109 (2015) 212–225.

[20] F. Ibarrola, R. Spies, A two-step mixed inpainting method with curvature-based anisotropy and spatial adaptivity, Inverse Problems & Imaging 11 (2).

[21] V. Peterson, H. L. Rufiner, R. D. Spies, Generalized sparse discriminant analysis for event-related potential classification, Biomedical Signal Processing and Control 35 (2017) 70–78.

[22] G. Mazzieri, R. Spies, K. Temperini, Mixed spatially varying L^2-BV regularization of inverse ill-posed problems, Journal of Inverse and Ill-posed Problems 23 (6) (2015) 571–585.

[23] Y. Avargel, I. Cohen, System identification in the short-time Fourier transform domain with crossband filtering, IEEE Transactions on Audio, Speech, and Language Processing 15 (4) (2007) 1305–1319.

[24] B. Yegnanarayana, P. S. Murthy, C. Avendaño, H. Hermansky, Enhancement of reverberant speech using LP residual, in: Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 1, IEEE, 1998, pp. 405–408.

[25] P. Smaragdis, Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs, Proceedings of the 5th Conference on Independent Component Analysis and Blind Signal Separation (2004) 494–499.

[26] C. Bouman, K. Sauer, A generalized gaussian image model for edge-preserving map estimation, IEEE Transactions on Image Processing 2 (3) (1993) 296–310.

[27] E. De Sena, N. Antonello, M. Moonen, T. Van Waterschoot, On the modeling of rectangular geometries in room acoustic simulations, IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 23 (4) (2015) 774–786.

[28] R. Ratnam, D. L. Jones, B. C. Wheeler, W. D. OBrien Jr, C. R. Lansing, A. S. Feng, Blind estimation of reverberation time, The Journal of the Acoustical Society of America 114 (5) (2003) 2877–2892.

[29] K. Matsuoka, Minimal distortion principle for blind source separation, in: 41st SICE Annual Conference, Vol. 4, IEEE, 2002, pp. 2138–2143.

[30] D. R. Hunter, K. Lange, A tutorial on MM algorithms, The American Statistician 58 (1) (2004) 30–37.

[31] D. D. Lee, H. S. Seung, Algorithms for non-negative matrix factorization, in: Advances in Neural Information Processing Systems, 2001, pp. 556–562.

[32] V. Zue, S. Seneff, J. Glass, Speech database development at MIT: TIMIT and beyond, Speech Communication 9 (4) (1990) 351–356.

[33] J. B. Allen, D. A. Berkley, Image method for efficiently simulating small-room acoustics, The Journal of the Acoustical Society of America 65 (4) (1979) 943–950.

[34] K. Kinoshita, M. Delcroix, S. Gannot, E. A. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, et al., A summary of the reverb challenge: state-of-the-art and remaining challenges in reverberant speech processing research, EURASIP Journal on Advances in Signal Processing 2016 (1) (2016) 7.

[35] Y. Hu, P. C. Loizou, Evaluation of objective quality measures for speech enhancement, IEEE Transactions on Audio, Speech, and Language Processing 16 (1) (2008) 229–238.

[36] T. H. Falk, C. Zheng, W.-Y. Chan, A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech, IEEE Transactions on Audio, Speech, and Language Processing 18 (7) (2010) 1766–1774.

[37] A. Farina, Advancements in impulse response measurements by sine sweeps, in: Audio Engineering Society Convention 122, Audio Engineering Society, 2007.